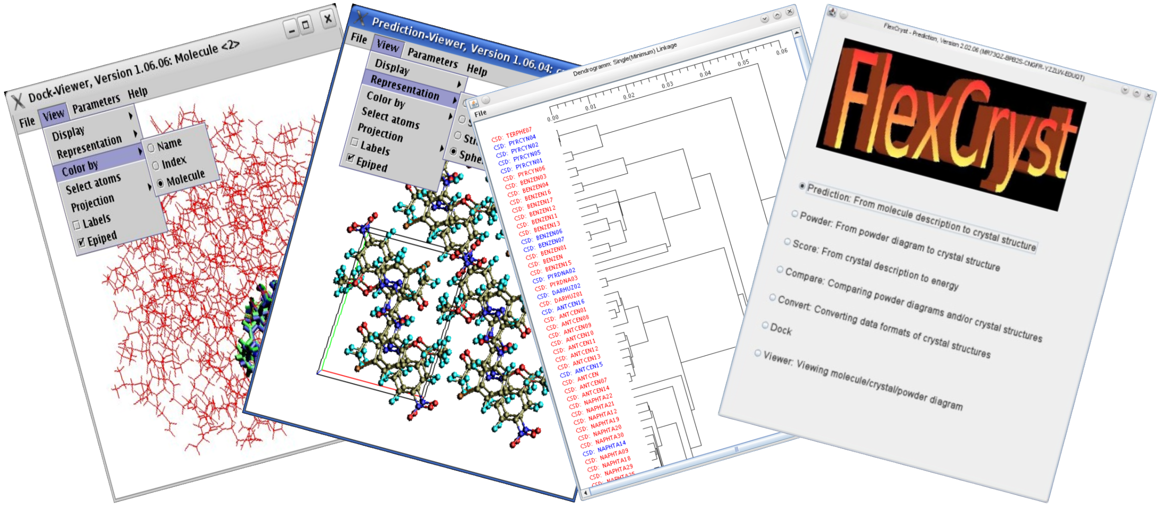
Prediction of Metal Coordination by Data Mining of experimental Crystal Structures (D.W.M. Hofmann)

CRS4
IDEAS BECOME LIFE

*Center for Advanced Studies,
Research and Development in
Sardinia (CRS4)
Loc. Piscina Manna, Edificio 1
09010 Pula
Italy*

http://www.crs4.it


Dock-Viewer, Version 1.06.06: Molecule <2>
File
View
Parameters
Help
Display
Representation
Color by
Select atoms
Projection
Labels
Epiped
Name
Index
Molecule
Prediction-Viewer, Version 1.06.04:
Dendrogram: Single(Minimum) Linkage
FlexCryst

○ Prediction: From molecule description to crystal structure
○ Powder: From powder diagram to crystal structure
○ Score: From crystal description to energy
○ Compare: Comparing powder diagrams and/or crystal structures
○ Convert: Converting data formats of crystal structures
○ Dock
○ Viewer: Viewing molecule/crystal/powder diagram

FlexCryst

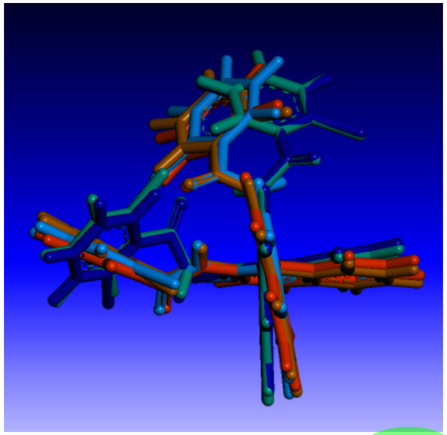
*FlexCryst
Schleifweg 23
91080 Uttenreuth
Germany*

http://www.flexcryst.com

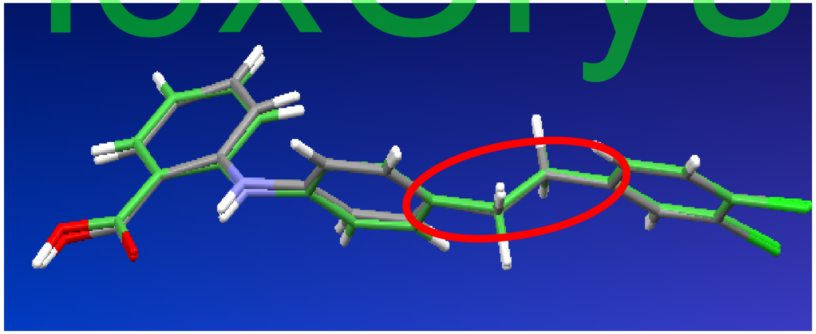
Definition of Atom Groups

FlexCryst did work in the past strictly with rigid molecules. They are several reasons to change this:

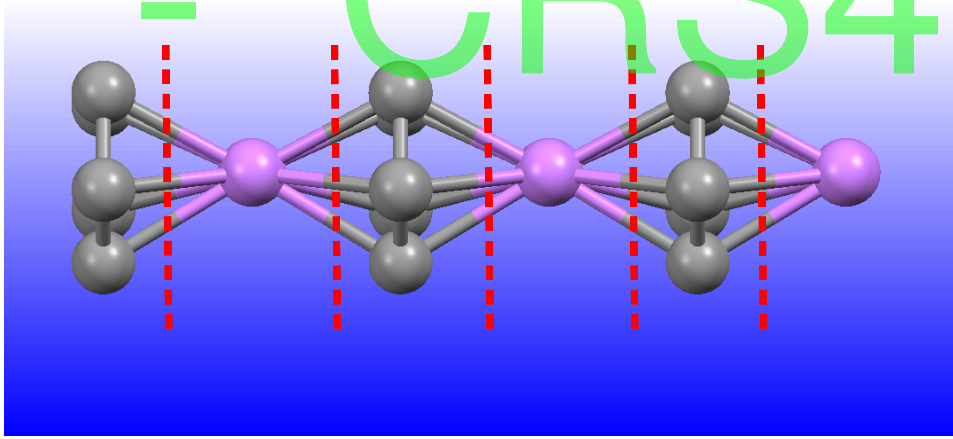
- Polymers are infinite, if they are not divided in finite moieties
- Small changes in the conformation have large effects in elongated molecules on distant moieties
- The number of reasonable conformations growth exponentially with number of degrees of freedom



The picture shows highly flexible molecule from the last blind test on crystal structure prediction at Cambridge. It has seven degrees of freedom, which is presently considered as a maximum for feasible structures. Assuming for any degree of freedom 4 possibilities, one obtains in total 4^7=16384 conformations.



The picture shows an elongated molecule from the last blind test on crystal structure prediction at Cambridge. Small errors in the indicated torsion angle fail a correct prediction of the crystal structure.

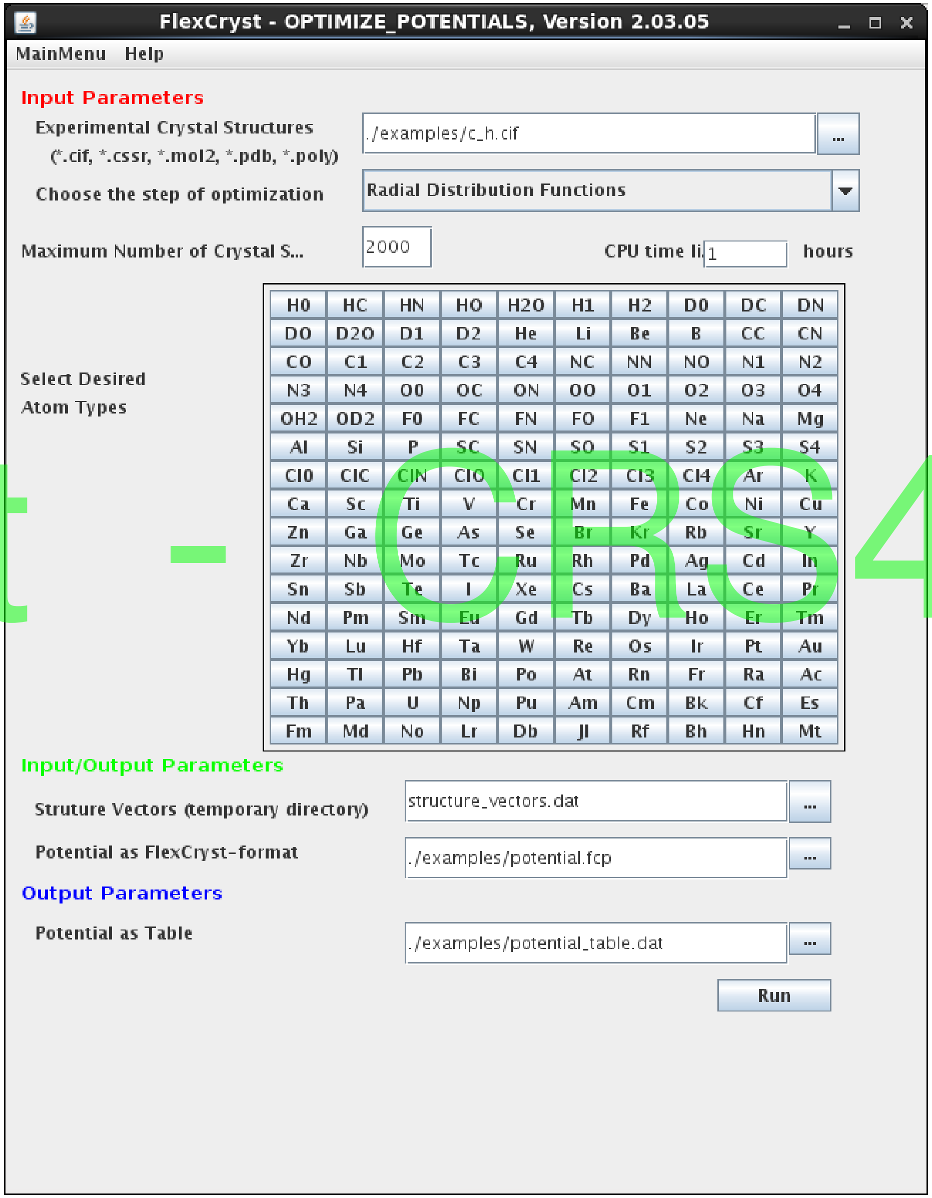


Li(Cp) serves as test, if the molecule can be divided in moieties and a Data Mining Force Field can be derived, which will correctly describe the .structure

Definition of the Atom Types

The atom types are assigned in systematic way and in case of multiple choice the more strong rule has priority. The rules are in the following order:

- If the atom belongs to a special molecule, it get its own type (presently this concerns water, H2O and OH2)
- If the atom is frequent (C, H, D, N, O, S, F, and Cl) and is bonded only to one other frequent atom, it get his own chemical symbol followed by the symbol for the second atom (e.g. NC for cyanid or ON for nitro compounds)
- If the atom is frequent, it get his own chemical symbol followed by the number of bonded atoms to it (e.g. C4 for sp3-carbon atoms or Cl0 for chlorid ions)
- In all other cases the atom type is identical to the chemical symbol



Input Parameters

Experimental Crystal Structures (*.cif, *.cssr, *.mol2, *.pdb, *.poly): ./examples/c_h.cif ...

Choose the step of optimization: Radial Distribution Functions

Maximum Number of Crystal S...: 2000

CPU time li: 1 hours

Select Desired Atom Types

H0	HC	HN	HO	H2O	H1	H2	D0	DC	DN
DO	D2O	D1	D2	He	Li	Be	B	CC	CN
CO	C1	C2	C3	C4	NC	NN	NO	N1	N2
N3	N4	O0	OC	ON	OO	O1	O2	O3	O4
OH2	OD2	F0	FC	FN	FO	F1	Ne	Na	Mg
Al	Si	P	SC	SN	SO	S1	S2	S3	S4
Cl0	ClC	ClN	ClO	Cl1	Cl2	Cl3	Cl4	Ar	K
Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu
Zn	Ga	Ge	As	Se	Br	Kr	Rb	Sr	Y
Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In
Sn	Sb	Te	I	Xe	Cs	Ba	La	Ce	Pr
Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm
Yb	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au
Hg	Tl	Pb	Bi	Po	At	Rn	Fr	Ra	Ac
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es
Fm	Md	No	Lr	Db	Jl	Rf	Bh	Hn	Mt

Input/Output Parameters

Struture Vectors (temporary directory): structure_vectors.dat ...

Potential as FlexCryst-format: ./examples/potential.fcp ...

Output Parameters

Potential as Table: ./examples/potential_table.dat ...

Effective Central Force Field

Commonly the force fields divide the interactions in intra- and intermolecular interactions. The intermolecular interactions are described by spherical functions, e.g. Lennard-Jones, Buckingham, etc., and the intramolecular interactions are described by energy functions for the bond distance, the bond angle, the torsion angles, etc. This breakdown of the problem is not furthermore feasible for organometallic and inorganic crystals. Simple rules for bond angles and torsion angles are non-existent for the coordination around metals.

Therefore we used the approach of effective spherical energy functions, known as central force field.The main features of central force field:

- The central force model was introduced for water (J. Chem. Phys. (1975) 62, 1677)
- The potentials reproduce experimental data rather than to provide a physical interpretation
- The molecular interactions do not broken down in inter- and intra-molecular interactions.
- The energy function might have several minima.

The interaction between two molecules I an J is divided in atom pair energies g(i,j,r):

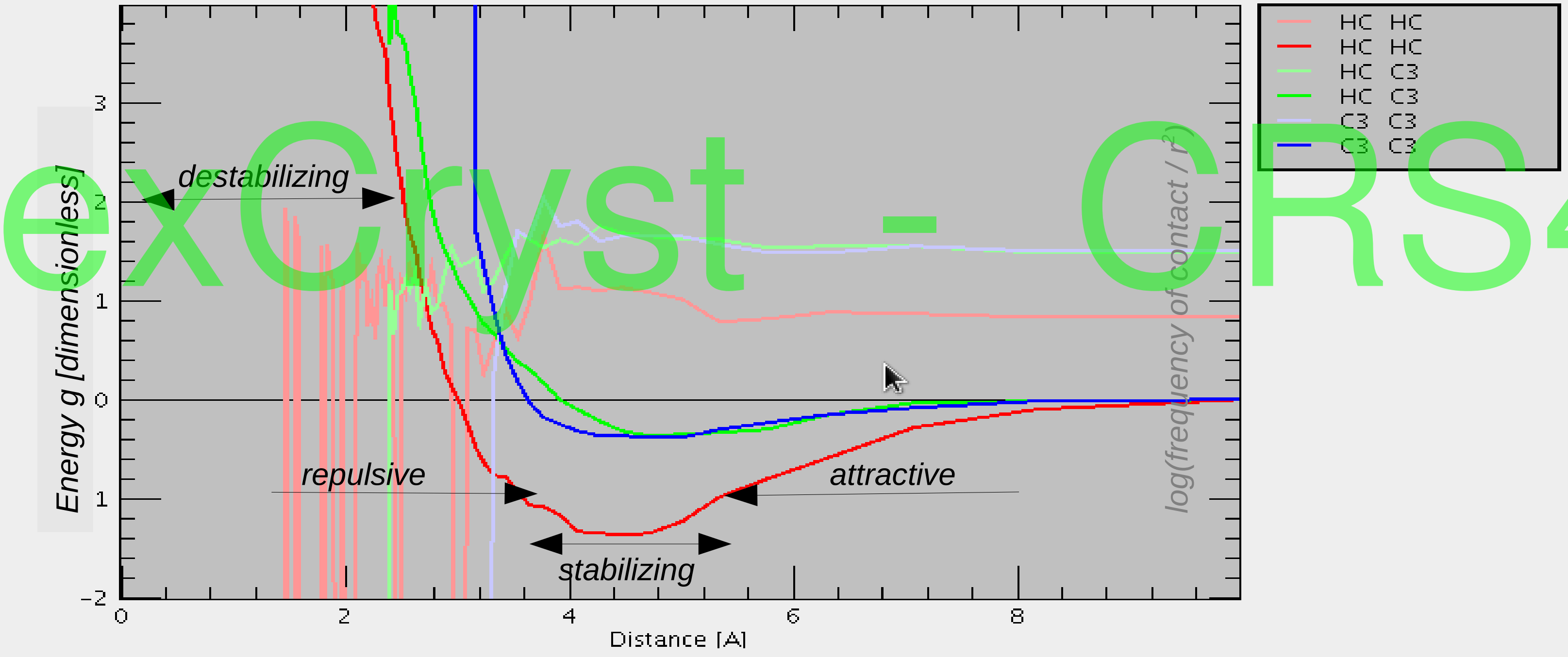
$$G_{IJ}=\sum_{i=1}^{N_I}\sum_{j=1}^{N_J} g(type_i, type_j, r_{ij})$$

Any force field can be transformed to a central force field by an series expansion, e.g by the inverse Taylor series:

$$g_{mnr}=\sum_{1=1}^{\infty} a_{mn}/(r^k)^i$$

Radial Distribution Function (RDF) and Energy

In the central force field the energy of a molecule can be easily visualized. The energy is the scalar product between the occurring distances and the energies of the relevant atom pair function. The sum of all occurring atom pair distances is commonly named pair correlation function in molecules and radial distribution function in liquids.



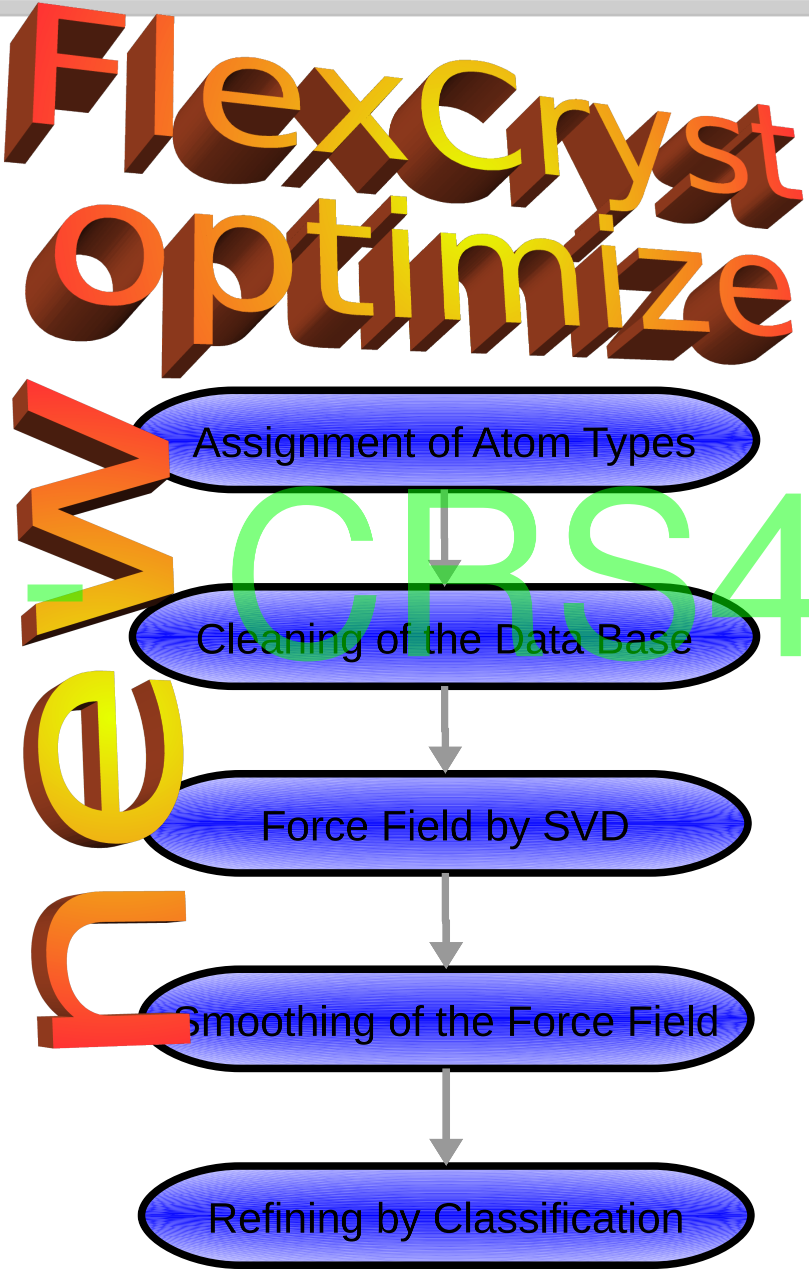
In this picture the energy calculation is demonstrated for crystal of BOXGAW. The energy is given as

$$G_{IJ}=\sum_{i=1}^{N_I}\sum_{j=1}^{N_J} g(type_i,type_j,r_{ij})=\vec{g}\cdot\vec{f}+\vec{g}\cdot\vec{f}+\vec{g}\cdot\vec{f}$$

Flowchart of the Force Field Derivation

In the past we observed three main problems for FlexCryst in the crystal structure prediction:

- ➢ The energy is not accurate enough. It is to simple, if it does not differentiate between atom types.
- ➢ The restriction to rigid molecules does not allow the prediction of large molecules. Even small changes in bond angles or torsion angles have a big effect on distant atom groups.
- ➢ The optimization procedure of the potentials depends strongly on the initial potential.



Assignment of Atom Types

Cleaning of the Data Base

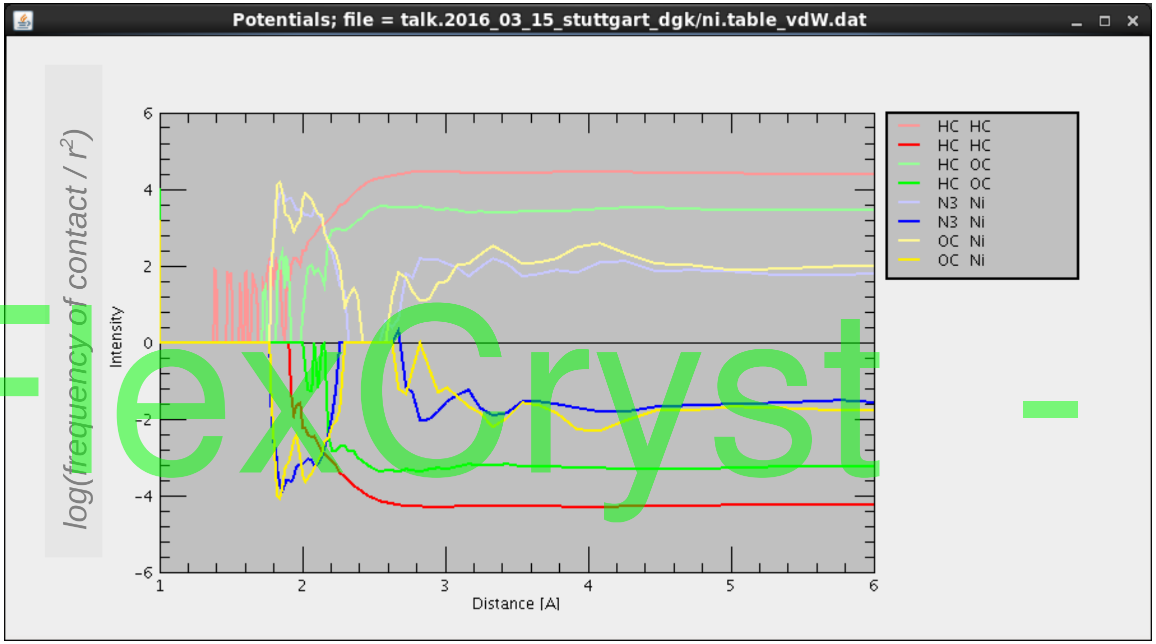
Force Field by SVD

Smoothing of the Force Field

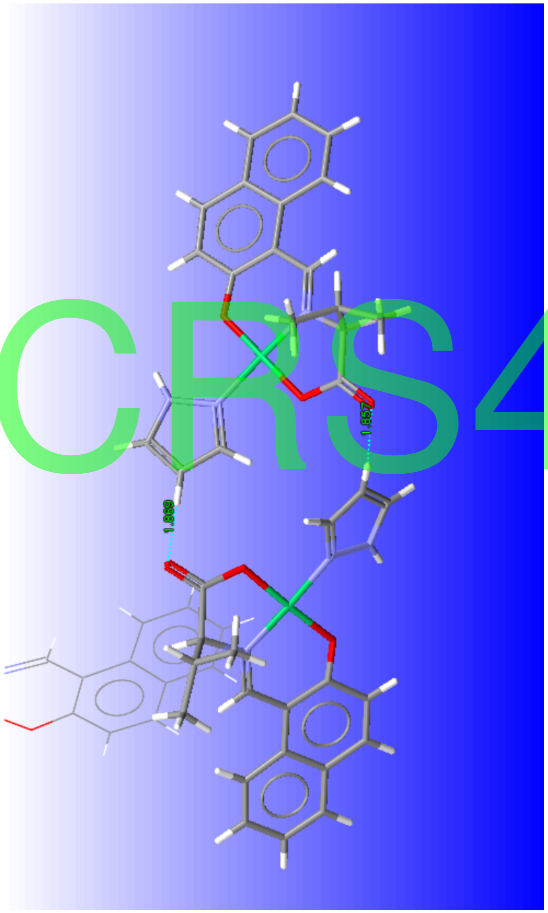
Refining by Classification

Anomaly Detection

A first step in data mining is always the cleaning of the data base by anomaly detection. For this purpose we plot all occurring atom distances and look for unusual distances. The structures with unusual distances are eliminated from the future process.



Potentials; file = talk.2016_03_15_stuttgart_dgk/ni.table_vdW.dat



In this figure we performed an analysis of the structure with FlexCryst. We plotted the frequency of the contacts for all Ni-compounds. While the peaks around 2 A for the radial distribution function of Ni reflect the nickel bonds to carbon and nitrogen, the RDF's of hydrogen-hydrogen and hydrogen–oxygen shows unusual contacts at short distances. We received for four structures a warning:

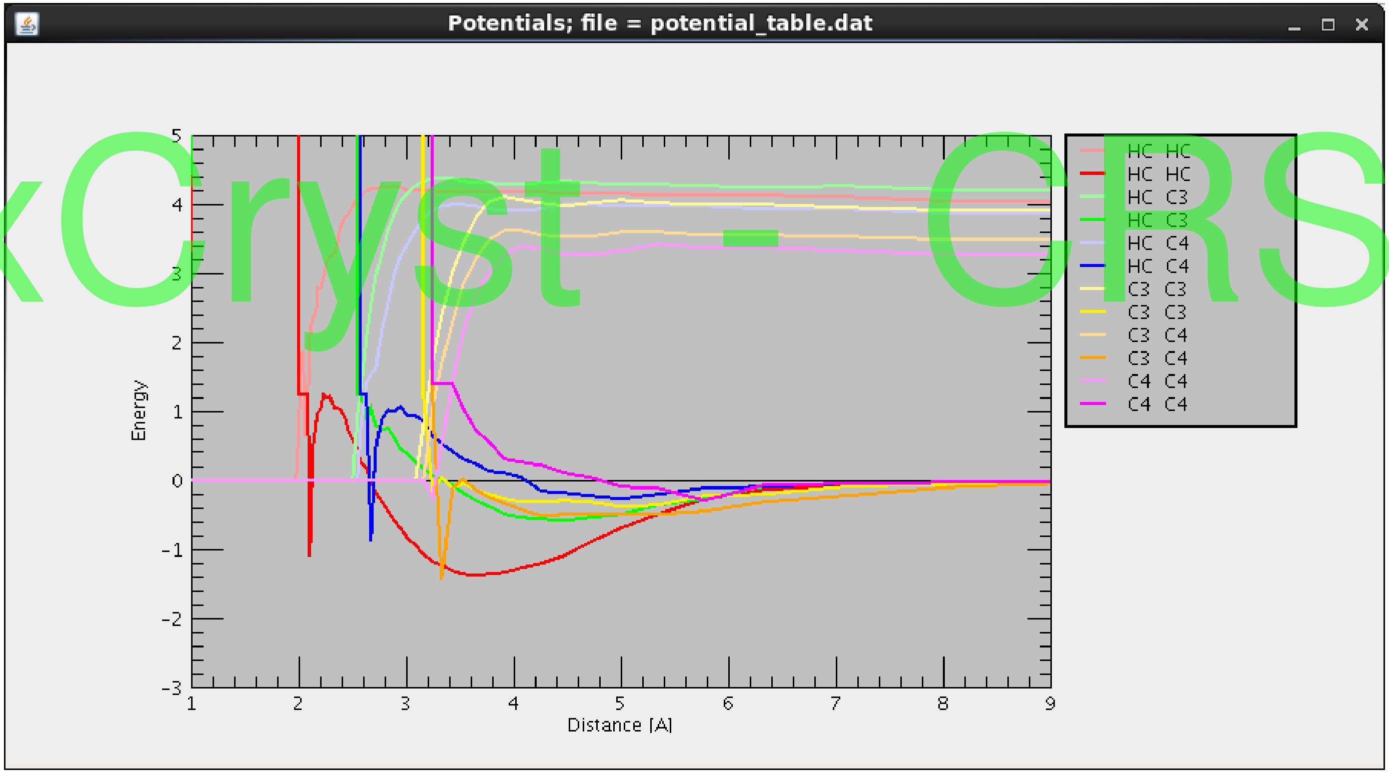
CSD: HIXJEF	*HC HC contact with a gap, d=1.82*
CSD: QOXCIP	*HC HC contact with a gap, d=1.78*
CSD: SUYBOE	*HC OC contact with a gap, d=1.85*
CSD: YEHMAZ	*HC HC contact with a gap, d=1.81*

The structure SUYBOE with an unusual hydrogen-oxygen contact

Force Field by Singular Value Decomposition

Any experimental crystal structure is a local minimum. This means that the gradient referring to the elongation or rotation of any molecule or atom must be zero. This gives us a large number of equations. If in addition the model is linearized by dividing the atom pair potentials of interest in piecewise linear sections, we obtain an overdetermined linear equations system. For this kind of systems an optimal solution can be obtained by singular value decomposition.

D.W.M. Hofmann & L. N. Kuleshova
International Tables of Crystallography C
in preparation



Potentials; file = potential_table.dat

The picture shows the radial distribution function and the obtained potentials by singular value decomposition. While the solution is the optimum from the mathematical point of view, it gives unreasonable solutions for rare atom distances and needs further improvement.

Refinement of the Force Field by Classification

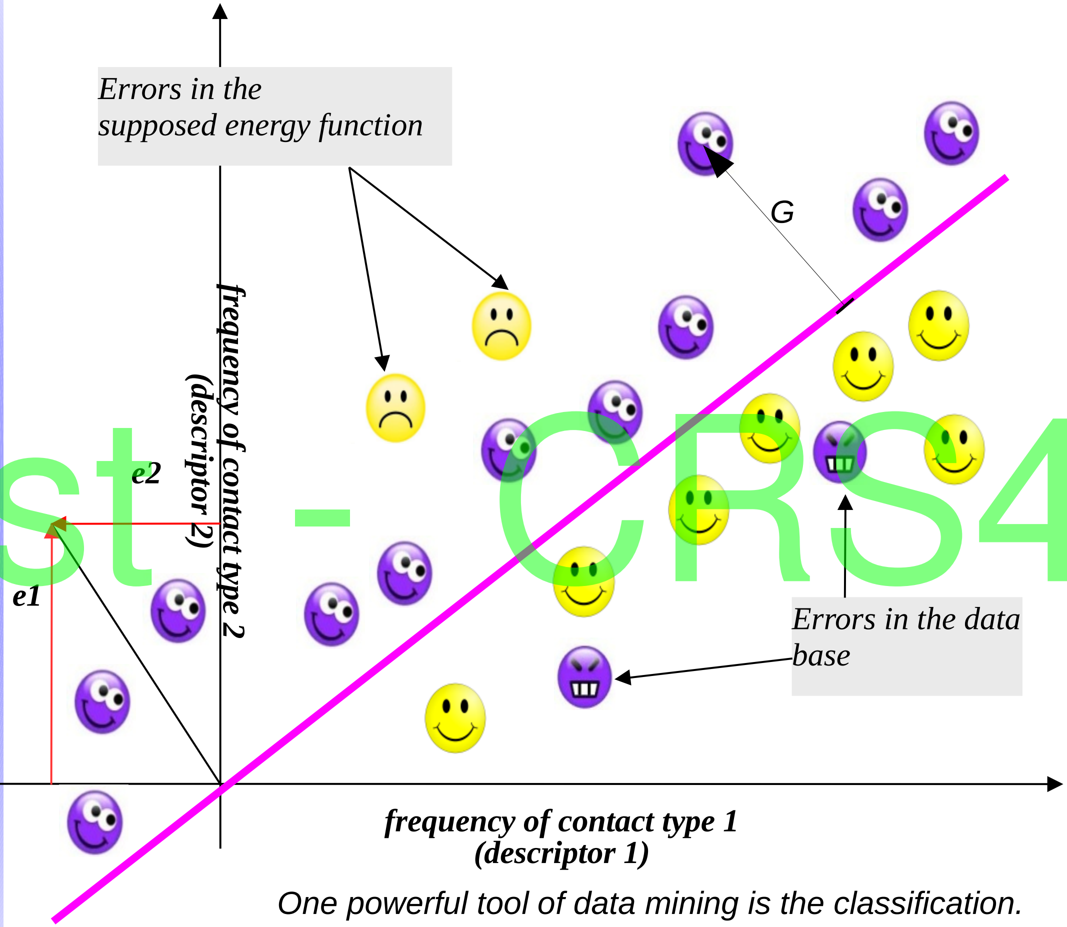
The parameters ***a*** of an arbitrary (energy) function can be determined by supervised learning. Aim of the minimization problem is to optimize the separation between the two classes and to keep the error for the outliers as small as possible. For any experimental structure the function should assign a negative energy and for any difference between experimental and virtual structure a positive energy.

$$G_k^{experimental} < 0$$

$$G_{km}^{virtual} - G_k^{experimental} > 0$$

$$G = \sum g_{mn}(r, \vec{a}) * frequency_{mnr}$$

For the solution of the problem can be used for instance a support vector machine.

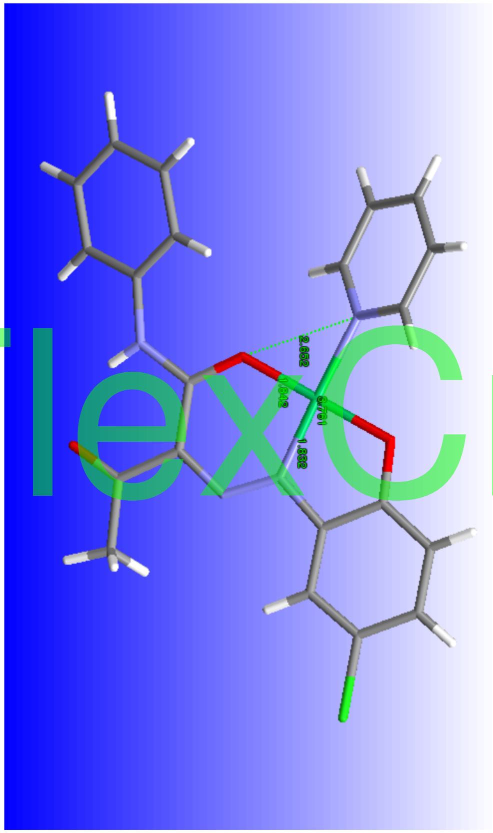

Errors in the supposed energy function
G
e2
e1
frequency of contact type 2 (descriptor 2)
Errors in the data base
frequency of contact type 1 (descriptor 1)

One powerful tool of data mining is the classification. By classification correct and incorrect crystal structures are separated. We can assume that experiment is always right and theory has always some (small) error. Aim of the classification is to divide by classification the structures in the two classes, experimental structures and predicted structures.

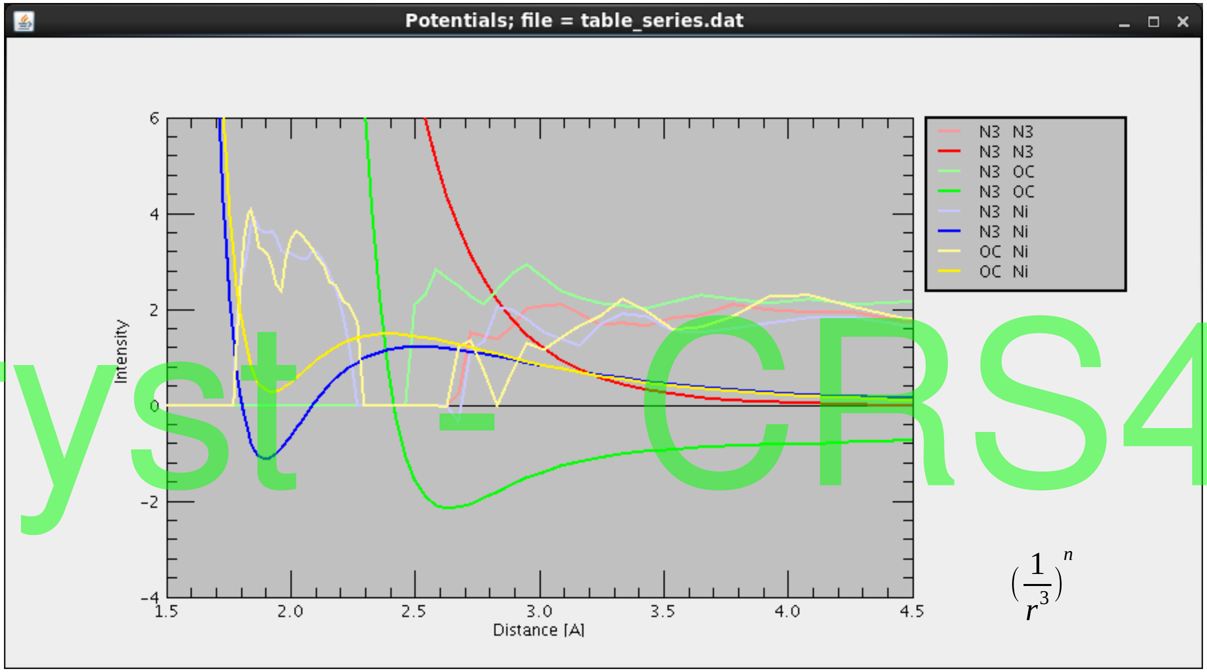
D.W.M. Hofmann
Data Mining in in Organic Crystallography
Structure and Bonding 134, Springer Verlag (2010)

Atom Pair Potential Functions

The obtained atom pair functions include the inter- and the intra-molecular interaction, including the metal bonds.



One of the structures studied in detail is the structure QEPSEK. It is a nickel-complex coordinated with a pyridine and tridental chelate ligand.



$$\left(\frac{1}{r^3}\right)^n$$

As result the optimization gives the force field

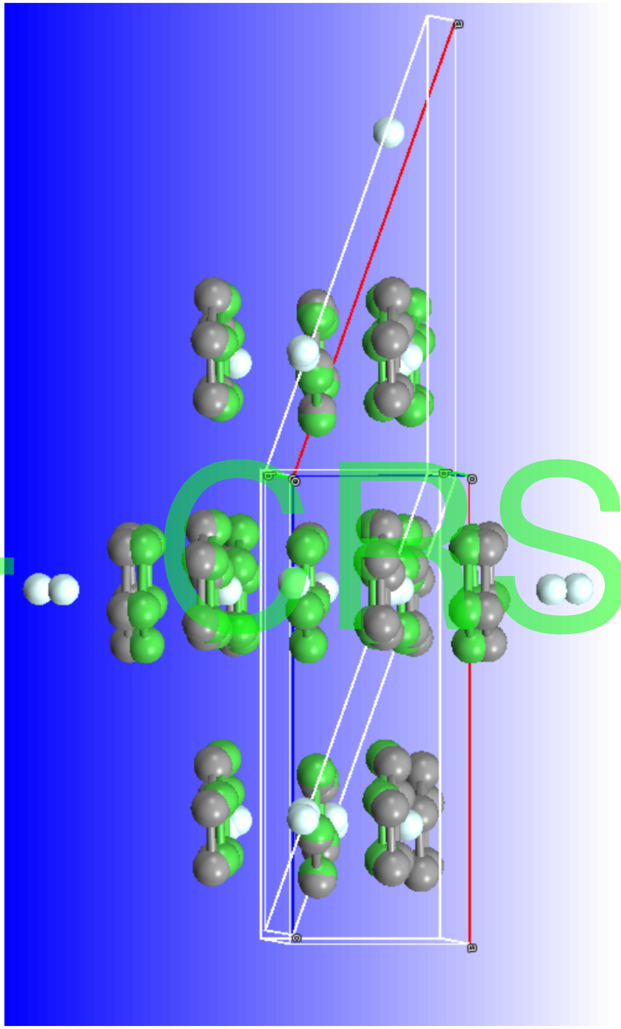
```
HC HC  0.00000    0.00000    -48.7769    0.00000    0.00000    2077.90    0.00000    0.00000    -8615.24    0.00000    0.00000    59906.6
HC HN  0.00000    0.00000    -18.7752    0.00000    0.00000    850.977    0.00000    0.00000    -7924.71    0.00000    0.00000    27008.6
HC HO  0.00000    0.00000    -28.2835    0.00000    0.00000    1304.60    0.00000    0.00000    -13959.1    0.00000    0.00000    63901.5
HC C2  0.00000    0.00000    51.1560     0.00000    0.00000    -1608.55   0.00000    0.00000    1434.66     0.00000    0.00000    255009.
HC C3  0.00000    0.00000    -29.9645    0.00000    0.00000    2160.24    0.00000    0.00000    -18434.0    0.00000    0.00000    130629.
HC C4  0.00000    0.00000    -49.9190    0.00000    0.00000    3885.00    0.00000    0.00000    -68015.4    0.00000    0.00000    403546.
...
```

1428 structures have been used to derive the potential. 48831 decoys have been generated for the training. 105 potentials with 388 have been derived. The mean error for any decoy 0.12 % of the total energy. 6.5% of the structures are outliers. The largest outlier contributes to 1.3 % to the total error.

Validation by Minimizing

The most simple way for the validation of a force field is the minimization of experimental structures with it. The structure should change only a few since the force field should have a local minimum for any experimental structure. We check the change in the density, in the energy, and in the overlay of the structure.

Reference code	BIWYUD	QEPSEK	KONBOE	NIBSEV
Estimated density	0.986	1.52	1.41	1.03
Experimental density	0.944	1.58	1.42	1.06
Minimized density	0.992	1.58	1.44	1.24
Error in the density	-4.94	-0.03%	-1.44%	-15.1%
Experimental energy	-251.91	-934.08	-1692.4	-489.39
Minimized energy	-270.97	-954.31	-1702.8	-505.13
Error in the energy	2.54%	1.95 %	0.53%	3.03%



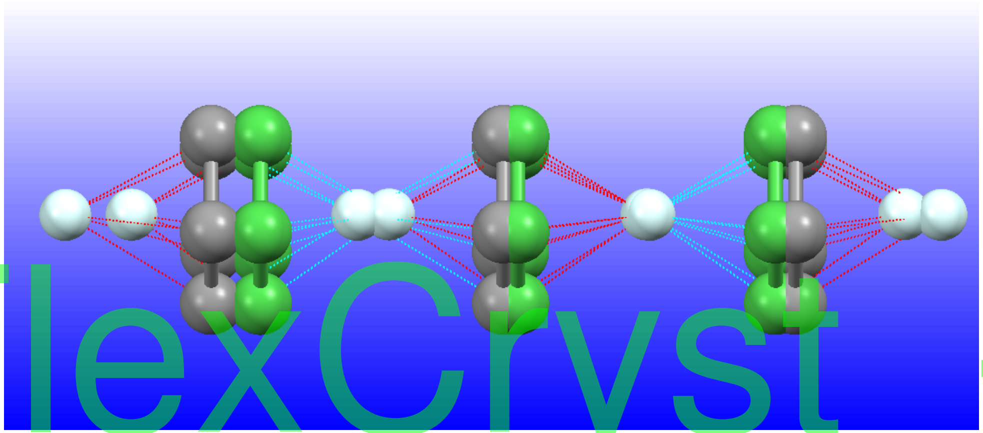
The relative error is defined slightly different as usual. In this way the absolute value is invariant against the exchange of a and b (necessary for clustering).

$$\eta = \frac{a-b}{\sqrt{(a^2+b^2)/2}}$$

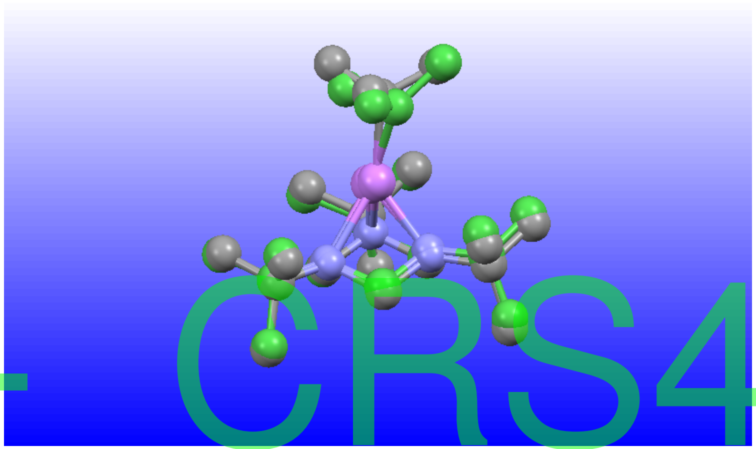
For the minimization the space group symmetry is lowered from Pnma to P21/a to avoid molecules on specific positions, For the visualization the Li has been replaced by He, since Mercury do not allow for polymers in the module "crystal packing similarity".

Prediction of the Coordination

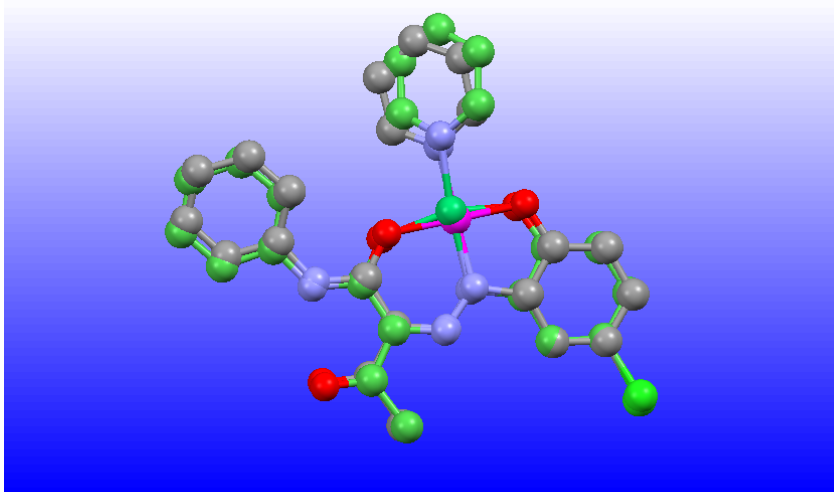
In all predictions we find the experimental structure between the first one hundred proposed structures. In simple cases as Li(Cp) the correct structure is ranked the first, in more complex structures this is presently not the case.



The picture shows the superposition of the experimental and the predicted coordination of the Li-atom for the crystal NIBSEW. All predicted structures have this coordination.

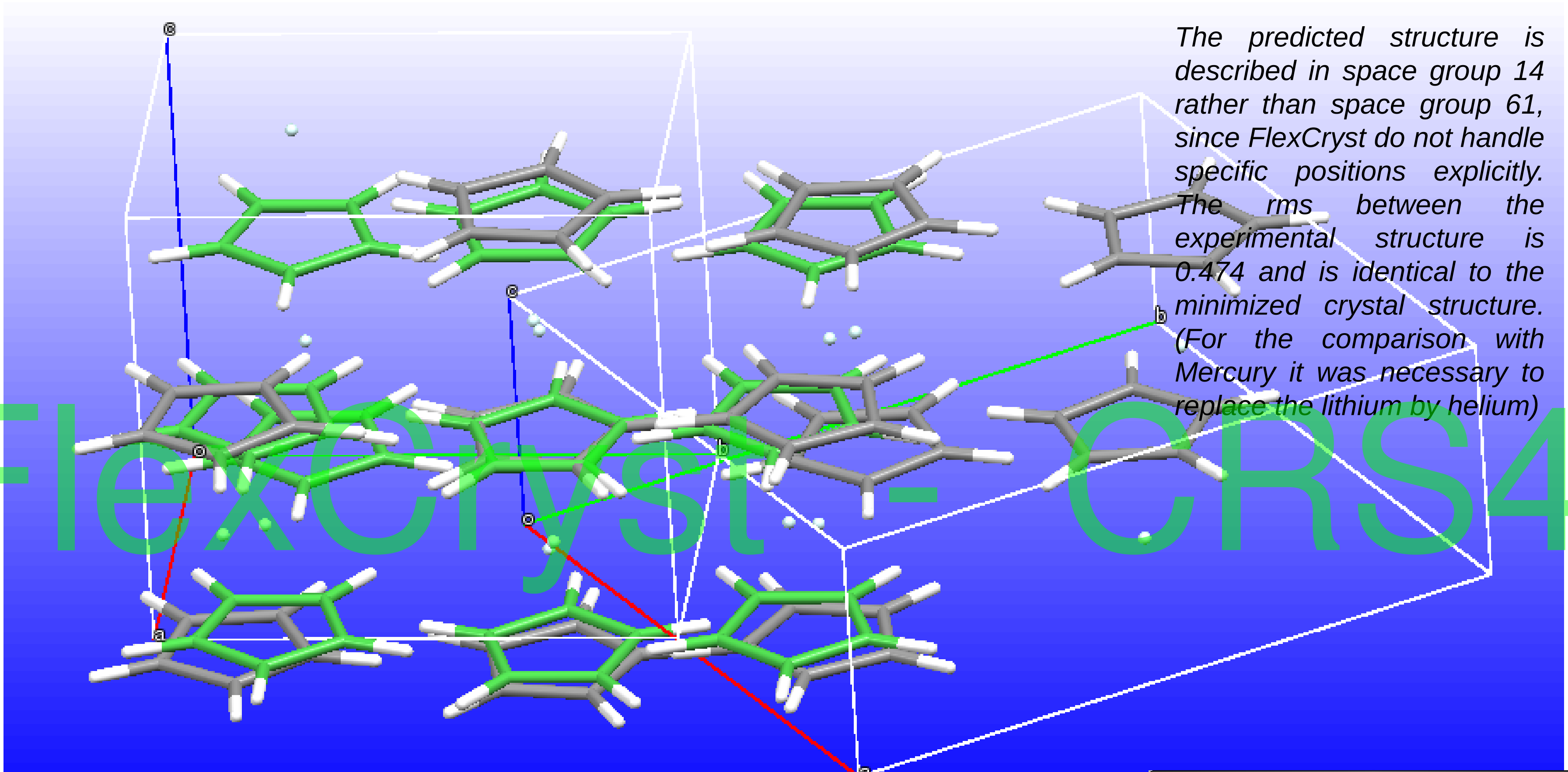


Another example is the structure BIWYUD. The tert-Butyllithium is stabilized by three coordination bonds. We find the experimental staggered configuration (rank 117) as well as the eclipsed conformation (rank 15).



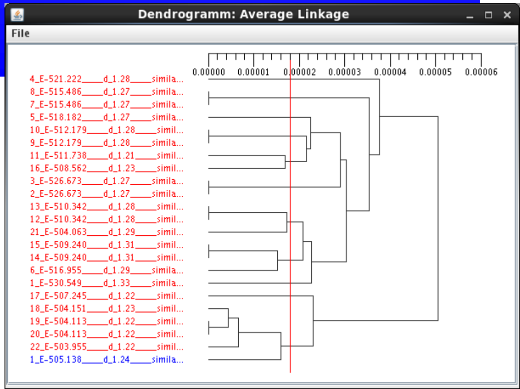
For the chelate-pyridin-nickel complex QEPSEK of nickel we find the experimental coordination on rank 6.

Prediction of the Crystal Structure



The predicted structure is described in space group 14 rather than space group 61, since FlexCryst do not handle specific positions explicitly. The rms between the experimental structure is 0.474 and is identical to the minimized crystal structure. (For the comparison with Mercury it was necessary to replace the lithium by helium)

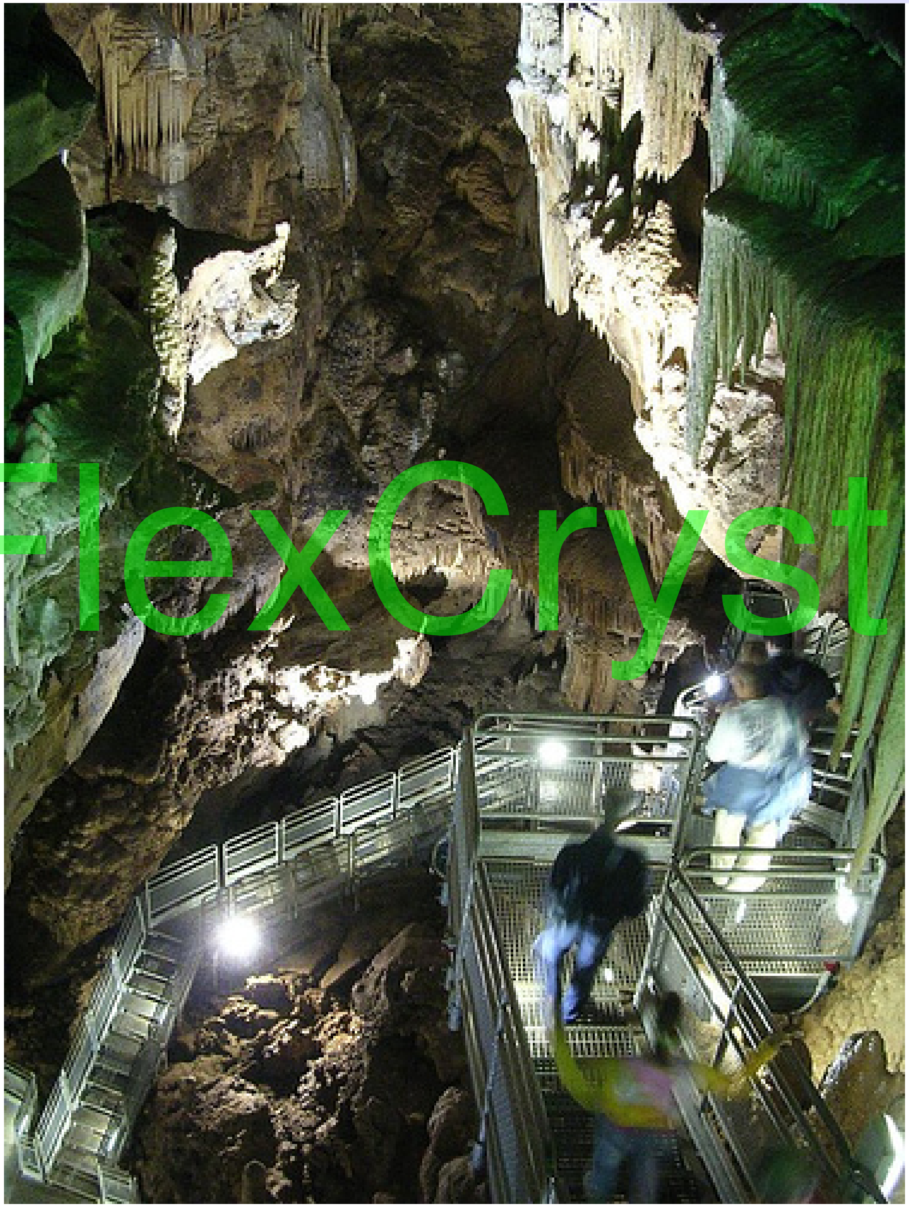
The prediction for the crystal structure of Li(Cp) gives the experimental structure four times on rank 18, 19, 20, and 22. This becomes obvious by a clustering of the simulated powder diagrams.



```
4_E-521.222_____d_1.28_____simila...
8_E-515.486_____d_1.27_____simila...
7_E-515.486_____d_1.27_____simila...
5_E-518.182_____d_1.27_____simila...
10_E-512.179_____d_1.28_____simil...
9_E-512.179_____d_1.28_____simila...
11_E-511.738_____d_1.21_____simil...
16_E-508.562_____d_1.23_____simil...
3_E-526.673_____d_1.27_____simila...
2_E-526.673_____d_1.27_____simila...
13_E-510.342_____d_1.28_____simil...
12_E-510.342_____d_1.28_____simil...
21_E-504.063_____d_1.29_____simil...
15_E-509.240_____d_1.31_____simil...
14_E-509.240_____d_1.31_____simil...
6_E-516.955_____d_1.29_____simila...
1_E-530.549_____d_1.33_____simila...
17_E-507.245_____d_1.22_____simil...
18_E-504.151_____d_1.23_____simil...
19_E-504.113_____d_1.22_____simil...
20_E-504.113_____d_1.22_____simil...
22_E-503.955_____d_1.22_____simil...
1_E-505.138_____d_1.24_____simila...
```

D.W.M.Hofmann and L.N.Kuleshova,
New similarity index for crystal structure determination from X-ray powder diagrams
J. Appl. Cryst. (2005) 38, 861.

Thanks for your Attention



Aragonite in the cave of Is Zuddas (Sardegna, Italy)

Calcite in the cave of Su Mannau (Sardegna, Italy)